DOCUMENT RESUME

ED 262 040 TM 850 488

AUTHOR Diamond, Esther E.
TITLE Content Considerations vs. Growth in Achievement Testing: Hobson's Choice?
PUB DATE Apr 84
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984)
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Gains; *Achievement Tests; Age Differences; *Difficulty Level; Elementary Secondary Education; Item Analysis; *Norm Referenced Tests; Pretesting; Standardized Tests; *Test Construction; Testing Problems; Test Items; *Test Validity; Trend Analysis
IDENTIFIERS Curricular Validity; *SRA Survey of Basic Skills; Test Batteries

ABSTRACT

The problem of measuring growth across the target grade and, typically, the two adjacent grades, concerns most developers of standardized, norm-referenced achievement tests, particularly at the item selection stage. Opinion is divided on whether to retain or drop items that do not get easier from grade to grade. The controversy has focused on whether test design should be guided mainly be test theory or by instructional content, with the latter viewpoint recently gaining strength. Trial test items from a new standardized achievement test battery, the survey of Basic Skills, were selected. The items covered a number of grade levels and content areas, seemed well constructed, showed good discrimination difficulty, but did not show grade-to-grade growth. These items were examined in an attempt to discover which kinds of items did not show grade-to-grade-growth; they were categorized according to curriculum content and grouped by grades 1-2, 2-3, 5-6, 8-10, and 11-12. It was concluded that the test developer's choice between measuring important content and showing growth from grade to grade is still difficult. In selecting the best available items, based on multiple criteria, one dimension is occasionally sacrificed in order to accommodate the other. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED262040

Content Considerations vs. Growth in Achievement Testing: Hobson's Choice?

Esther E. Diamond
Educational and Psychological Consultant
Evanston, Illinois

Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, April 1984

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
E. E. Diamond
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 850 488

Content Considerations vs. Growth in Achievement Testing:

Hobson's Choice?

Abstract

The problem of achieving growth across the target grade and, typically, the two adjacent grades, is a vexing one that plagues most developers of standardized, norm-referenced achievement tests, particularly at the item selection stage, working with raw data from the pretest. Opinion is somewhat divided on whether to retain or drop items that do not get easer from grade to grade. The controversy has centered largely around the question of whether test design should be guided mainly by test score theory or by instructional or cognitive theory (i.e., content concerns), with the latter viewpoint gaining strength in recent years. Data for pretest items from a standardized achievement test battery are discussed, as is the dilemma for test publishers, who select what appear to be the best items based on multiple criteria and must sometimes sacrifice one dimension in order to accommodate the other.

Content Considerations vs. Growth in Achievement Testing: Hobson's Choice?

When test publishers develop an achievement test battery, they typically start out with an analysis--usually by curriculum content experts--of the common elements in school curricula across the country. The task is to determine, for each major subject area, both the specific curricular objectives and the longer-range educational behavior objectives at each of the grade levels for which the test is intended. A subject-matter x objectives grid is then constructed, and relative weights assigned to each cell in the grid. Items are then written to sample randomly from each domain in the cell. For pretest purposes, about three times as many items may be written as are planned for inclusion in the final test.

Pretest items are then tried out with samples from the target population and, generally, with samples from adjacent grades as well, in order to determine if the item shows grade-to-grade growth in the skill or knowledge being tested. For example, a test designed primarily for use in the fifth grade may be tried out with samples of fourth- and sixth-graders as well. If, for example, 40% of the tryout sample in grade 4, 50% in grade 5, and 57% in grade 6 get an item right, the item is considered to demonstrate grade-to-grade growth and--providing difficulty and discrimination indices are within prescribed limits--is included in the test. If, however, 50% of the grade 4 sample, 48% of the grade 5 sample, and 45% of the grade 6 sample get it right, the test developer faces a decision that can be critical--particularly if there are not very many items surviving the pretest that measure the same objective but have more favorable growth statistics.

The Content/Growth Controversy

Current opinion is somewhat divided on whether to retain or drop items that measure important curriculum content or objectives but do not get easier

from grade to grade. The controversy has centered largely around the question of whether test design should be guided mainly by test score theory or by instructional or cognitive theory (i.e., content concerns)(Linn, 1980). A test designed for the purpose of measuring achievement, Linn argued, should place its primary emphasis on item content: "The measurement of achievement may or may not result in good differentiation among students and it may or may not show rapid or steady growth from grade to grade" (p. 84). The goal in measuring achievement, he stated, quoting Glaser (1963), is to provide explicit information on what a student can and cannot do.

Giving preference in item selection to items that show the largest decrease in difficulty from grade to grade, Linn pointed out further, "may distort the content validity of an achievement test" for the items remaining in the test after the "no-growth" items have been eliminated "are not necessarily representative of the content domain that the test is designed to measure" (1980, p. 84).

A similar position was taken earlier by Buros (1978), who argued that the practice of eliminating items that did not show grade-to-grade growth concealed "unlearning" and gave "an illusory sense of continuous learning from grade to grade" (p. 1975). Normative information, he suggested, should include information on whether each objective and each item was taught in an earlier grade, taught in the target grade, taught in later grades, or not taught at all. To illustrate his point, Buros cited his own experience as principal and test director in a small school system. American history was taught in the fifth grade and European backgrounds in the sixth, and he expected sixth-graders to do less well in American history, when tested a year later, because of forgetting. Actually, they did better, because approximately one-third of the pretest items, which had not shown sufficient grade-to-grade growth, had been eliminated.

Green (1983),on the other hand, held that "objectives that do not show growth through the grades can be included in criterion-referenced tests but not in a test which must provide unambiguous norms." If the material taught becomes part of the student's structure of understanding or operational skills, Green maintained, performance is not likely to decline, whereas material independent of such skills will tend to be forgotten by many. Does this mean, then, that much of the important body of knowledge in science, mathematics, and history, for example, should not be measured in an achievement test because students may have forgotten much of the material by the time they move on to higher grades?

As we know from classical learning theory (e.g., Morgan, 1956), the acquisition of many skills is not a one-stage process but often involves several different stages. Learners may reach a plateau when they complete one stage of learning and before the next stage begins. Morgan gives learning to play the piano as an illustration: Fundamental skills are quickly acquired, and then there is a long period of little or no improvement. Forgetting, retroactive inhibition of what has been learned, and temporary loss of motivation may also play a part. In educational achievement, the fact that there is no additional teaching in the subject area in subsequent grades may account for the failure to show growth. However, rarely is material learned totally forgotten, and skills that have become rusty will generally be relearned in far less time than it took to learn them initially. Measurement of these skills is important, even if they do not show growth. They may be essential to a course of study a student decides to pursue in later school years.

Ebel (1969), perhaps one of the strongest proponents of emphasis on subject matter content in achievement testing, challenged use of "the obvious facts of forgetting" as an objection to emphasis on knowledge in schools. "...(A)n item of useful information that is well integrated into a structure of knowledge is not likely to be forgotten," he stated. "Much of the information a person receives is quickly, and often quite appropriately, forgotten. But the knowledge that a person succeeds in developing, on the basis of his old knowledge and his new information, tends to be much more permanent" (pp. 71-72).

The subject matter or content dimension of two-way-grid test blueprints, Ebel maintained, was far more important than the abilities or objectives dimension for the purpose of achieving "an adequate sampling of the most useful knowledge" (p. 73). Of the structure of knowledge, he stated:

> Every factual detail and every generalization that can be related to other factual details and generalizations becomes part of the structure. The function of this structure...is to give the facts some degree of coherence and thus to make them meaningful and useful...What the student finds in any good textbook of physics or economics or history is such a structure. (p. 69)

Illustrations from an Achievement Battery

In the item selection process for a new achievement battery--the Survey of Basic Skills (SBS), soon to be published by Science Research Associates--there was an opportunity to examine in depth a variety of items at a number of grade levels and in a spread of content areas that met every requirement except that of grade-to-grade growth for the pretest samples.

Sometimes it was possible to offer logical explanations, but often there seemed to be no apparent reason for the lack of growth. The items were generally well constructed and unambiguous, they did not test trivia, and they showed good discrimination and appropriate difficulty--and even before they were sacrificed, based on just the data from the pretest samples, the subtests of which they were a part showed average positive gainsin raw scores from grade to grade. Since subtests cover a variety of skills and objectives, this is not surprising; enough items and the objectives they measured showed sufficient growth to more than compensate for those that did not.

Which kinds of items, then, showed growth and which did not? The illustrations given here are not intended to be comprehensive, but merely to provide a broad overview of the problem across subtests at selected grade levels. Table 1 summarizes this information. (Math items were not reviewed by this author.)

Insert Table 1 about here

Grades 1, 2, and 3

Of items tried out in grades 1 and 2, in Auditory Discrimination only a few--testing different beginning sound(s) and different middle or vowel sounds--showed no growth. It is possible that it may take more than one grade for this skill to become part of a child's repertoire of skills. Letters and sounds, on the other hand, showed marked grade-to-grade growth, with identification of the ch sound in chair increasing in p value from .48 to .84, while others showed increases from .55 to .86 and from .71 to .97. In Listening Comprehension, a few items on understanding directions and on grasping details showed no growth or--in one or two cases--a slight decline.

By grade 3, Letters and Sounds showed good growth throughout, with the exception of one item on final consonants. In LIstening Comprehension, some literal items (understanding directions, grasping details) failed to show growth, as did several inferential items (main idea, cause and effect, inferred idea, character/emotion), although--on the whole--inferential items showed greater growth from grade 2 to 3 than did literal items.

Most of the Vocabulary items exhibited strong growth from grade 2 to grade 3, but a few showed a decline. No logical explanation could be found. Reading Comprehension showed strong growth throughout; all of the items increased in p value. In Mechanics, three skill areas exhibited growth problems, although they were confined to only a handful of items--first-letter alphabetization, use of the question mark, and use of the exclamation point. The nine other skill areas showed good growth for all items.

Grades 5 and 6

In English Usage in the middle grades (5 and 6), growth problems showed up in some items dealing with possessive nouns and antecedent pronouns, although possessive pronouns were problem-free. There were no problems, either, in items that tested verb agreement, modifiers, coordinating conjunctions, and prepositions. One item that dealt with present perfect tense showed no growth. In writing skills, an item testing understanding of topic sentences showed no growth, as was the case with a few sentence-recognition items, but all of the other writing skills items showed good grade-to-grade growth, with sentence order and recognition of irrelevant sentences showing the strongest growth of all.

In Social Studies, growth problems were found mainly in items measuring recall objectives in geography, history, and sociology. Again, forgetting

may have played a major role. Interpretation and reasoning problems generally did well.

In Science, growth problems were found with several items on photosynthesis, astronomy, sedimentation, genetics, and food chains. An item about current (electromagnetism) displayed the greatest growth problems, items about geophysics and science philosophy the least. Items about plant and animal structures and systems performed best of all. (It is interesting to note, however, that at the next higher level plant structures and functions items had more growth problems, while there were none among items on photosynthesis, green plants vs. mold, and physical states, and almost none in items on mechanics, heat, and electromagnetism. Items on observation and proof, scientific hypotheses, laws, and theories also did well as far as growth was concerned.)

In Spelling, the main problem area was the combination of vowels or consonants with vowels, although in later grades some items testing suffixes and variant pronunciation showed no growth. Also, a number of words customarily introduced at one level and retaught at others continued to pose problems--for example, ie and ei words.

Grades 8, 9, and 10

At this level, gains in Reading Comprehension items were very slight from grade 8 to grade 9 and moderate from grade 9 to grade 10. Most students in the sample may have reached a plateau in reading comprehension skills by grade 9 or 10. Sizeable gains in fact/opinion and cause/effect items from grade 9 to grade 10 were offset by losses, later, in grade 11. Other items that showed no gain from grade 9 to grade 10 involved author's purpose, drawing conclusions, and figurative language.

Spelling exhibited very few growth problems, but several Mechanics items involving the various uses of quotation marks and capitalization showed declines or no growth from grade 10 to 11. In Usage, the problem areas were collective and plural nouns, possessive pronouns, negative modifiers, and relative pronouns.

In Reference Materials, although there was growth across grades 8-10 for the subtest as a whole, more items showed decline from grade 8 to 9 than from grade 9 to 10. These involved use of bibliographies, indexes, card catalogs, and the parts of a book.

In Social Studies, map- and graph-reading skills showed some grade-to-grade plateauing or decline. There were relatively few no-growth items in Science; one item involved use of scientific tests for photosynthesis, and another the reading of graphic data.

Grades 11 and 12

At this level, items that showed no growth or actual decline were very similar to problem items at the next lower level, and they can be summed up quickly here. Usage items involved past tense, collective nouns, and recognition of complete sentences, paragraph reorganization, relative pronouns, and conjunctive adverbs. The few problem items in Reading Comprehension involved sequence, grasping details, recognizing emotions, drawing conclusions, and inferring the author's opinion. In Vocabulary, literal synonyms and nonliteral figurative language items exhibited occasional growth problems. In Social Studies, a few items on map and chart interpretation showed no growth or actual decline. In Science, an item on recognition or recall of scientific laws and another on scientific facts showed no growth, and one on comprehension of scientific law showed a slight decline. Mechanics problems, like those at the grades just below, involved capitalization and the use of quotes.

A Question Still Unanswered

Table 2 gives pretest raw-score means for the various subtests in selected grades. These data include scores on all items pretested. As noted earlier, even with no-growth items included, in all cases mean scores increased across grades. Note, however, that they did not increase in a steady progression. Letters and Sounds, for example, showed the greatest increase from grade 1 to grade 2 --possibly due, in part, to ceiling effect-- with more moderate increases for Concepts and Decoding, and smaller increases for the remaining subtests. Math Concepts, Computation, and Mechanics showed large increases from grade 2 to 3, and slight to moderate increases for the remaining subtests. Increases for all subtests were mostly moderate from grade 5 to grade 6. As we reach grades 8, 9, and 10, grade-to-grade increases are mostly slight--sometimes only a fraction of a raw-score point. Plateaus in certain skills apparently have been reached by this time.

Insert Table 2 about here

Comparable data for the standardization sample (comparison of average p values across grades) are not yet available, but they should show even greater grade-to-grade growth since, wherever possible, no-growth items were eliminated. The norms, consequently, will be--to use Green's (1983) term--"unambiguous" and will present no problems for test users. Pre-post test results and gain scores, as well as growth curves, normally will show favorable growth as students move from grade to grade, as is true for almost all standardized achievement tests today, no matter what the measurement mode they use--percentiles, grade equivalents, vertical scaled scores, and so on.

The question posed in this paper, however, remains unanswered: Should content, and the measurement of important knowledge, be sacrificed to maintain what Buros (1978) referred to as "an illusory sense of continuous

learning from grade to grade" (p. 1975)? The jury has been out on that question for a long, long time, and it is still a long way from reaching a decision. For test publishers, the choice between content and growth is realistically a Hobson's choice--selecting what appear to be the best available items, based on multiple criteria, and occasionally sacrificing one dimension in order to accommodate the other.

References

Buros, O.K. (1978). Fifty years in testing: Some reminiscenses, criticisms, and suggestions. In O.K. Buros (Ed.),The Eighth Mental Measurements Yearbook, Vol. 2 (pp. (pp. 1972-1979). Highland Park, N.J.: The Gryphon Press.

Ebel, R.L. (1969). Knowledge vs. ability in achievement testing. In Educational Testing Service, Toward a theory of achievement measurement, Proceedings of the 1969 Invitational Conference on Testing Problems (pp. 66-76). Princeton, N.J.: Educational Testing Service.

Green, D.R. Content validity of standardized achievement tests and test curriculum overlap. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, April 1983.

Linn, R.L. (1980). Test design and analysis for measurement of educational achievement. In W.B. Schrader(Ed.), Measuring achievement: Progress over a decade. New Directions for Testing and Measurement series, No. 5 (pp. 81-92). San Francisco: Jossey-Bass.

Morgan, C.T. Introduction to psychology. New York: McGraw-Hill, 1956.

Table 1

Types of Pretest Items Showing No Grade-to-Grade Growth

Grades	Subtest
1-2	Auditory Discrimination
	Different beginning sound or sounds
	Letters and Sounds
	Vowel digraphs
	Final consonant blends
	Decoding
	Beginning letters
	Listening Comprehension
	Understanding directions
	Grasping details
2-3	Letters and Sounds
	Final consonants
	Listening Comprehension
	Understanding directions
	Grasping details
	Inferring main idea, cause/effect, character/motive/emotion
	Vocabulary
	Literal synonyms
	Reading Comprehension
	None
	Mechanics
	Alphabetizing by first letter
	Question mark
	Exclamation point

table continues

Grades	Subtest
5-6	Vocabulary
	Nonliteral meanings
	Figurative language
	Reading Comprehension
	Sequence
	Title (for passage)
	Inferring character/motive/emotion
	Figures of speech
	Mechanics
	Capitalization: direct quote
	Apostrophe: contractions, possessives
	Coordinating conjunction
	Usage
	Possessive nouns
	Antecedent pronouns
	Present perfect tense
	Topic sentence
	Sentence recognition
	Spelling
	Vowel combinations
	Vowel/consonant combination
	Reference materials
	Card catalog
	Science
	Recall, recognition of scientific facts, categories, concepts
	Recall, application of appropriate scientific procedures

table continues

Grades	Subtest
5-6	Social Studies
	Recall of geography, history, and sociology facts
8-9-10	Vocabulary
	Literal meaning: synonyms
	Nonliteral meanings: figures of speech, idioms
	Antonyms
	Reading Comprehension
	Sequence
	Outcome
	Cause/effect
	Figurative language
	Author's purpose
	Mechanics
	Quotation marks
	Commas with quote
	Capitalization: proper noun, book title
	Usage
	Collective and plural nouns
	Possessive pronouns
	Relative pronouns
	Negative modifiers
	Spelling
	Affixes
	Contractions
	Reference Materials
	Bibliographies
	Indexes

table continues

Grades	Subtest
8-9-10	Reference Materials (continued)
	Card catalog
	Parts of books
	Science
	Reading graphic data
	Application of scientific tests
	Social Studies
	Map-reading
	Graph-reading
11-12	Vocabulary
	Literal meaning: synonyms
	Nonliteral meaning: idioms, figurative language
	Reading Comprehension
	Sequence
	Drawing conclusions
	Inferring emotions
	Understanding technique/style
	Mechanics
	Capitalization:proper noun, family name
	Direct quote
	Comma: direct address
	Punctuation, varied
	Usage
	Plural and collective nouns
	Past tense
	Recognizing complete sentence

table continues

Grades	Subtest
11-12	Usage (continued)
	Paragraph reorganization
	Negative modifier
	Relative pronoun
	Conjunctive adverb
	Spelling
	Affixes
	Consonants
	Commonly mispronounced words
	Reference Materials
	Graph-reading
	Map-reading
	Timetable (train)
	Bank statement
	Parts of book
	Card catalog
	Social Studies
	Map interpretation
	Chart reading
	Diagram interpretation
	Passage interpretation
	Science
	Use of scientific tests
	Recognition, comprehension of scientific laws
	Recall of scientific facts

Table 2

Pretest Raw-Score Means on SBS Subtests Across Selected Grades

Subtest	No. Items per Pretest Unit	Criterion Mean (Raw Score) Grade 1 (N=1000)	Grade 2 (N=440)
Letters and Sounds	35	13.09	19.18
Auditory Discrimination	23	14.18	16.93
Listening Comprehension	35	15.59	18.54
Concepts	57	17.56	21.51
Decoding	23	15.93	19.78
		Grade 2 (N=700)	Grade 3 (N= 530)
Letters and Sounds	35	19.76	22.55
Listening Comprehension	35	19.37	21.61
Vocabulary	55	24.15	27.75
Reading Comprehension	65	19.27	24.82
Concepts (Math)	57	28.86	36.33
Computation	57	29.07	37.26
Mechanics	45	22.73	31.64
		Grade 5 (N=500)	Grade 6 (N=500)
Vocabulary	55	34.14	39.30
Reading Comprehension	65	36.97	41.69
Concepts (Math)	57	23.43	27.26
Computation	57	29.44	35.43
Problem Solving	45	23.75	27.79
Mechanics	45	23.24	24.94
Usage	55	30.10	34.25

table continues

Subtest	No. Items per Pretest Unit	Criterion Mean (Raw Score) Grade 5 (continued)		Grade 6
Spelling	45	25.52		29.72
Reference Materials	55	34.21		39.44
Science	55	22.85		25.80
Social Studies	54	23.23		27.03
		Grade 8 (N=360)	Grade 9 (N=450)	Grade 10 (N=230)
Vocabulary	55	32.82	33.65	37.54
Reading Comprehension	65	37.76	38.40	41.93
Concepts (Math)	57	24.54	26.08	28.85
Computation	57	28.93	31.57	34.85
Problem Solving	45	23.24	23.88	26.84
Mechanics	45	23.90	24.45	24.97
Usage	55	30.86	32.02	33.32
Spelling	45	21.62	22.73	24.11
Reference Materials	55	33.79	34.51	36.39
Science	55	24.47	25.64	27.58
Social Studies	54	25.93	27.37	29.40